1. Show that in a simple linear regression model, F statistic for the F test for regression and t statistic for the test of $H_0: \beta_1 = 0$ are equivalent, that is, $F = t^2$. Show this by using the formulas for the test statistics.

For t-test, the formula for the test statistics is

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}, \quad SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \quad \hat{\sigma}^2 \text{ is } MSE.$$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. So $\hat{y}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$, $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$

So $(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 (x_i - \bar{x})^2$

For F-test, the formula for the test statistics is

$$F = \frac{MSReg}{MSE} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{MSE} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_1^2}{SE^2(\hat{\beta}_1)}$$

So $F = t^2$

2. Consider the simple linear regression model with normal errors, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim$ independent $N(0, \sigma^2)$, $i = 1, \dots, n$. A **linear transformation of a variable** w involves replacing w with a new variable $w^* = d + cw$, where c and d are constants. Explain how the quantities $\hat\beta_0$, $\hat\beta_1, \hat\sigma^2, R^2$ and the test of $H_0\colon \beta_1 = 0$ are affected by the following linear transformations. Provide mathematical justification for your answers.

   a. Each value of the predictor x_i is replaced by cx_i, where c is a non-zero constant. For example, the data set has age in months and we convert this to age in years by dividing by 12.
   b. Each value of x_i is replaced by $x_i + d$. For example, we subtract the mean $\bar{x}$ from each x_i.
   c. Each value of the response y_i is replaced by ky_i, for a non-zero constant k. For example, suppose that we convert income from units of \$1 to units of \$1000, by dividing by 1000.
   d. Each value of y_i is replaced by $y_i + d$. For example, we subtract the mean $\bar{y}$ from each y_i.
   e. Verify your answers to (a)-(d) by fitting models that regress HEIGHT on AGE in the SPIROMETRY data set. Fit models that: convert age in months to age in years; subtract mean age; convert height in cm to height in inches; subtract mean height. Compare to the model using the original scaling of the variables.

$\hat\beta_1 = \dfrac{\sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^n (x_i-\bar x)^2}$, $\hat\beta_0 = \bar y - \hat\beta_1 \bar x$, $\hat\sigma^2 = \dfrac{\sum_{i=1}^n (y_i - \hat y_i)^2}{n-2}$, $R^2 = \dfrac{\sum_{i=1}^n (\hat y_i - \bar y)^2}{\sum_{i=1}^n (y_i-\bar y)^2}$, test score $t = \dfrac{\hat\beta_1}{\frac{\hat\sigma}{\sqrt{\sum (x_i-\bar x)^2}}}$

a. $\hat\beta_1' = \dfrac{\sum_{i=1}^n c(x_i-\bar x)(y_i-\bar y)}{c^2 \sum_{i=1}^n (x_i-\bar x)^2} = \dfrac{1}{c}\hat\beta_1$ test score $t' = \dfrac{\hat\beta_1'}{\frac{\hat\sigma'}{\sqrt{\sum (x_i'-\bar x')^2}}} = \frac{1}{c}\hat\beta_1 \, c\sqrt{\sum_{i=1}^n (x_i-\bar x)^2}\,\frac{1}{\hat\sigma} = t$

$\hat\beta_0' = \bar y - \hat\beta_1' \bar x' = \bar y - \frac{1}{c}\hat\beta_1 c\bar x = \hat\beta_0$

$\hat y_i' = \hat\beta_1' x_i' + \hat\beta_0' = \frac{1}{c}\hat\beta_1 c x_i + \hat\beta_0 = \hat y_i$, so

$\hat\sigma^{2\prime} = \hat\sigma^2$ and $R^{2\prime} = R^2$

So $\hat\beta_1$ will become $\frac{1}{c}\hat\beta_1$. $\hat\beta_0$, $\hat\sigma^2$, R^2, test result will not change.

b. $x_i' = x_i + d$, $\bar x' = \bar x + d$, so $x_i' - \bar x' = x_i - \bar x$.

So $\hat\beta_1' = \hat\beta_1$

$\hat\beta_0' = \bar y - \hat\beta_1' \bar x' = \bar y - \hat\beta_1(\bar x + d) = \hat\beta_0 - \hat\beta_1 d$

$\hat y_i' = \hat\beta_1' x_i' + \hat\beta_0' = \hat\beta_1(x_i + d) + \hat\beta_0 - \hat\beta_1 d = \hat\beta_1 x_i + \hat\beta_0 = \hat y_i$

So $\hat\sigma^{2\prime} = \hat\sigma^2$ and $R^{2\prime} = R^2$. Since $\hat\beta_1$, $\hat\sigma$ and $x_i - \bar x$ do not change, then test score t does not change.

So $\hat\beta_1$, $\hat\sigma^2$, R^2, test result will not change. $\hat\beta_0$ will become $\hat\beta_0 - \hat\beta_1 d$.

c. $y_i' = k y_i$, $\bar y' = k\bar y$. $\hat\beta_1' = k\hat\beta_1$. $\hat\beta_0' = k\bar y - k\hat\beta_1 \bar x = k\hat\beta_0$

$\hat y_i' = \hat\beta_0' + \hat\beta_1' x_i = k\hat\beta_0 + k\hat\beta_1 x_i = k\hat y_i$ test score $t' = \dfrac{k\hat\beta_1}{\frac{k\hat\sigma}{\sqrt{\sum_{i=1}^n (x_i-\bar x)^2}}} = t$

So $\hat\sigma^{2\prime} = k^2\hat\sigma^2$. $R^{2\prime} = \dfrac{k^2 \sum_{i=1}^n (\hat y_i - \bar y)^2}{k^2 \sum_{i=1}^n (y_i - \bar y)^2} = R^2$

So $\hat\beta_1$ will become $k\hat\beta_1$, $\hat\beta_0$ will become $k\hat\beta_0$, $\hat\sigma^2$ will become $k^2\hat\sigma^2$. R^2 and test score will not change.

d. $y_i' = y_i + d$, $\bar y' = \bar y + d$, $y_i' - \bar y' = y_i - \bar y$. So $\hat\beta_1' = \hat\beta_1$.

$\hat\beta_0' = \bar y' - \hat\beta_1' \bar x = \bar y + d - \hat\beta_1 \bar x = \hat\beta_0 + d$

$\hat y_i' = \hat\beta_0' + \hat\beta_1' x_i = \hat\beta_0 + d + \hat\beta_1' x_i = \hat y_i + d$

So $y_i' - \hat y_i' = y_i + d - \hat y_i - d = y_i - \hat y_i$. $\hat y_i' - \bar y' = \hat y_i - \bar y$

So $\hat\sigma^{2\prime} = \hat\sigma^2$. $R^{2\prime} = R^2$. Since $\hat\beta_1$ and $\hat\sigma$ do not change, then test score t does not change.

So $\hat\beta_1$, $\hat\sigma^2$, R^2, test result will not change. $\hat\beta_0$ will become $\hat\beta_0 + d$.

e. Original model:

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

$\bar x = 53.53055$

$\bar y = 104.61972$

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	75.92256	1.49942	50.63	<.0001
AGE	Age (Months)	1	0.53609	0.02626	20.42	<.0001

model (a):

$\hat\beta_1' = 12 \times 0.53609 = 6.44308$

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	75.92256	1.49942	50.63	<.0001
AGE_in_year		1	6.43307	0.31507	20.42	<.0001

So $\hat\beta_1$ will become $12\hat\beta_1$. $\hat\beta_0$, $\hat\sigma^2$, R^2, test result will not change.

model (b):

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	104.61972	0.52235	200.29	<.0001
AGE_sub		1	0.53609	0.02626	20.42	<.0001

So $\hat\beta_1$, $\hat\sigma^2$, R^2 and test result will not change. $\hat\beta_0$ will become $\hat\beta_0 + \hat\beta_1 \bar x = \bar y$.

model (c):

Root MSE	1.73283	R-Square	0.8580
Dependent Mean	41.18887	Adj R-Sq	0.8559
Coeff Var	4.20702		

$\hat\beta_1' = \frac{1}{2.54}\hat\beta_1 = 0.21106$

$\hat\beta_0' = \frac{1}{2.54}\hat\beta_0 = 29.89077$

$\hat\sigma' = \frac{1}{2.54}\hat\sigma = 1.7328$, so $\hat\sigma^{2\prime} = \left(\frac{1}{2.54}\right)^2 \hat\sigma^2$.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	29.89077	0.59032	50.63	<.0001
AGE	Age (Months)	1	0.21106	0.01034	20.42	<.0001

So $\hat\beta_1, \hat\beta_0, \hat\sigma$ will become $\frac{1}{2.54}$ their original values. R^2 and test result will not change.

model (d):

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	9.859153E-9	Adj R-Sq	0.8559
Coeff Var	44642538851		

$\hat\beta_0' = -\hat\beta_1 \bar x = -0.53609 \times 53.53055 = -28.6972$

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-28.69716	1.49942	-19.14	<.0001
AGE	Age (Months)	1	0.53609	0.02626	20.42	<.0001

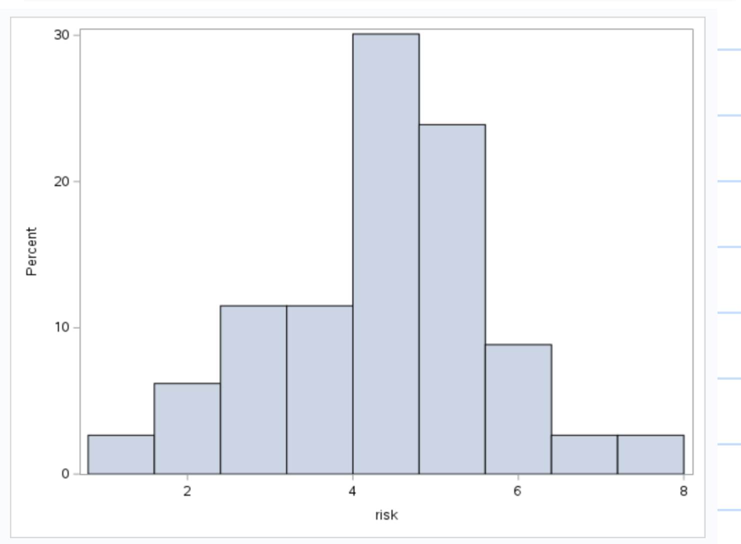
So $\hat\beta_1$, $\hat\sigma^2$, R^2 and test result will not change. $\hat\beta_0$ will become $\hat\beta_0 - \bar y = -\hat\beta_1 \bar x$

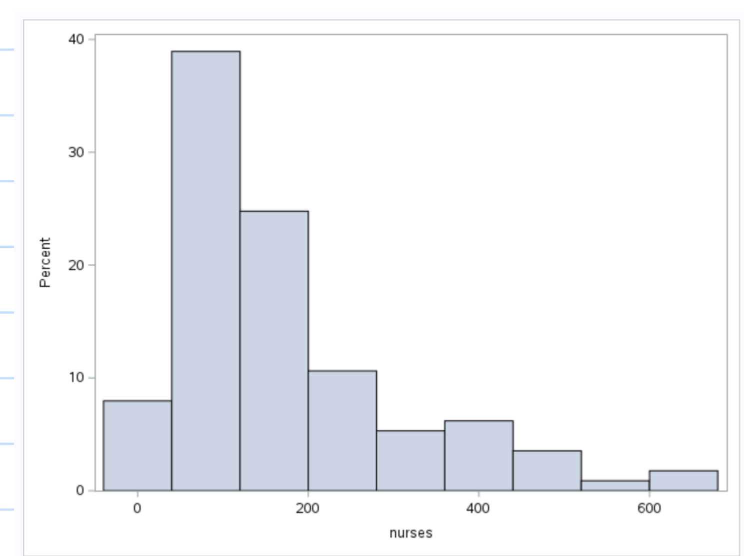
3. For the SENIC data, consider the variables risk, nurses, length and svcs. Obtain summary statistics and univariate plots of these variables (e.g., histograms) and describe briefly. Construct scatterplots of the following pairs of variables to examine their relationships (first variable is Y, second is X) and fit **loess curves**:
   a. Risk and nurses
   b. Risk and length
   c. Nurses and svcs

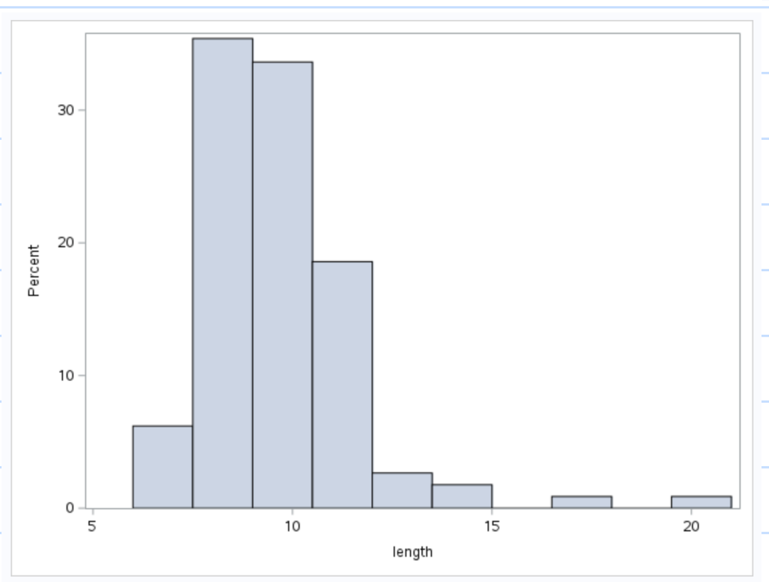
If the relationship in a scatterplot is nonlinear, attempt to make the relationship linear by transforming one or both variables, or explain why a **power/root transformation** to linearity is not an appropriate strategy. For each pair, are you able to find transformations such that the assumptions of the normal error regression model are approximately met? Conduct residuals analysis to determine this. What violations seem to remain and be difficult to address, if any? *Note: Often, the default level of smoothing in a software routine is too low; be sure to try different levels of smoothing.*

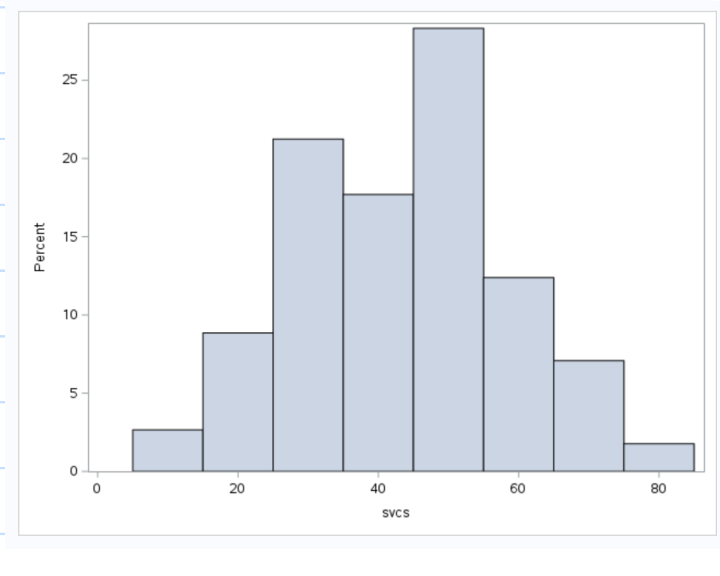
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
risk	113	4.3548673	1.3409080	1.3000000	7.8000002
nurses	113	173.2477876	139.2653897	14.0000000	656.0000000
length	113	9.6483186	1.9114560	6.6999998	19.5599995
svcs	113	43.1592918	15.2008613	5.6999998	80.0000000





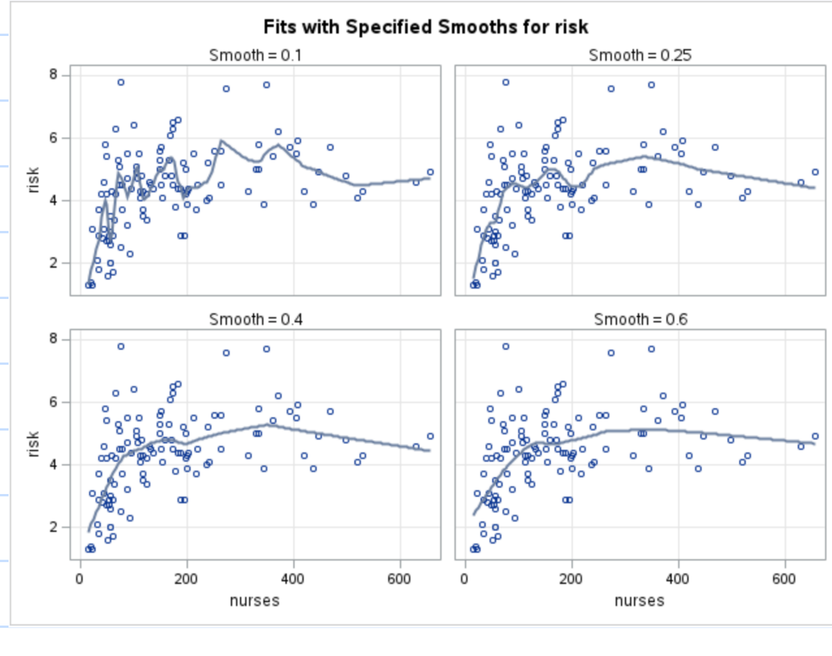


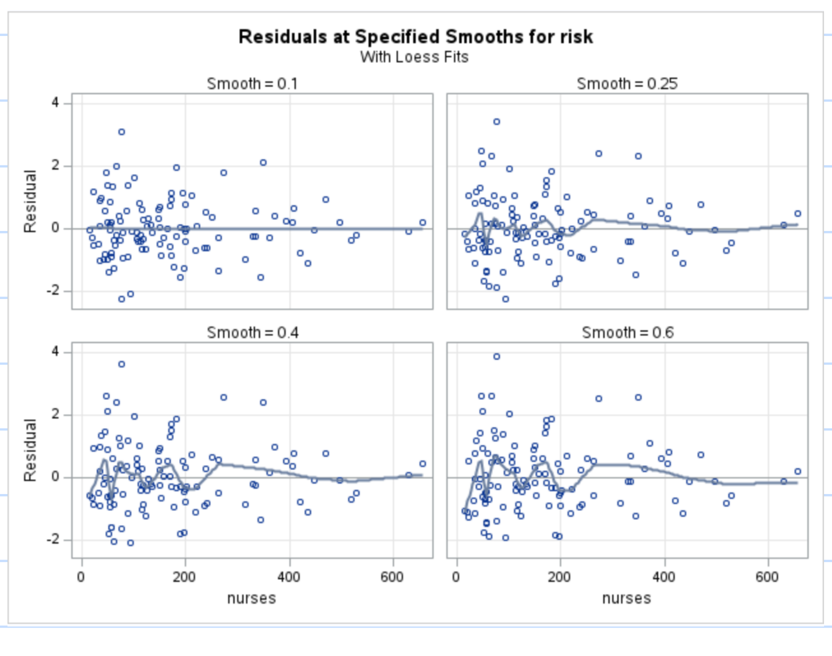


From the table, the mean of *risk* is 4.3548673 and the standard deviation is 1.3409080. The mean of *nurses* is 173.2477876 and the standard deviation is 139.2653897. The mean of *length* is 9.6483186 and the standard deviation is 1.9114560. The mean of *svcs* is 43.1592918 and the standard deviation is 15.2008613.

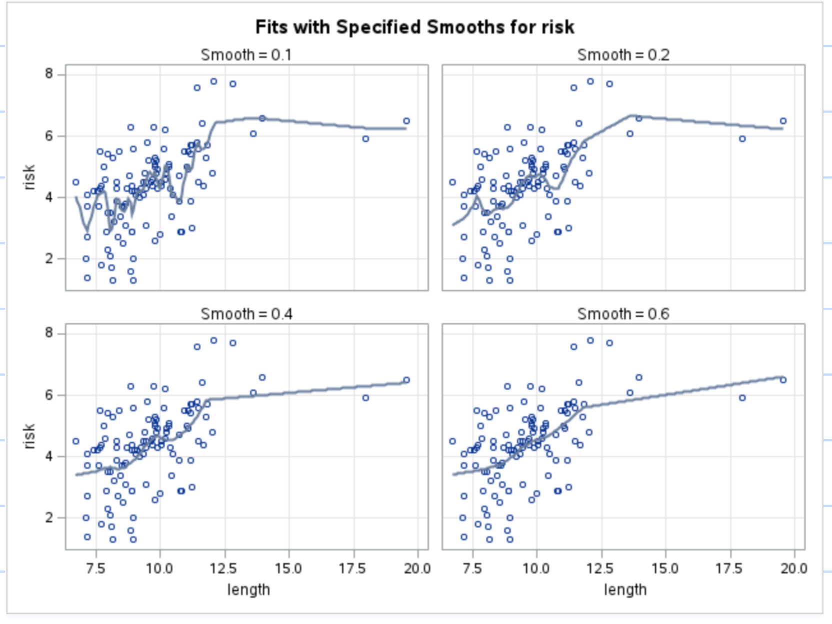
From their histograms, the distribution of *risk* is left-skewed. The distribution of *nurses* and *length* is right-skewed. The distribution of *svcs* is a normal distribution.

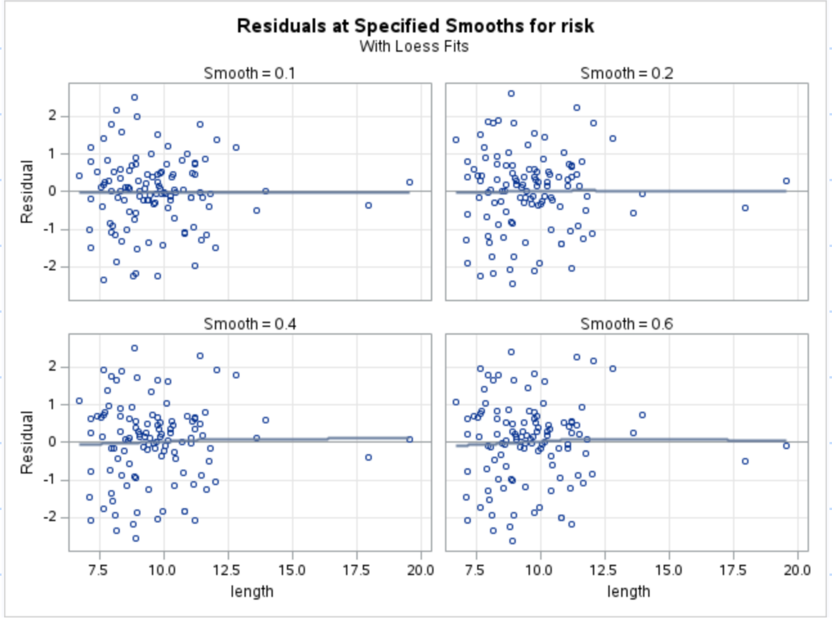
Model a:



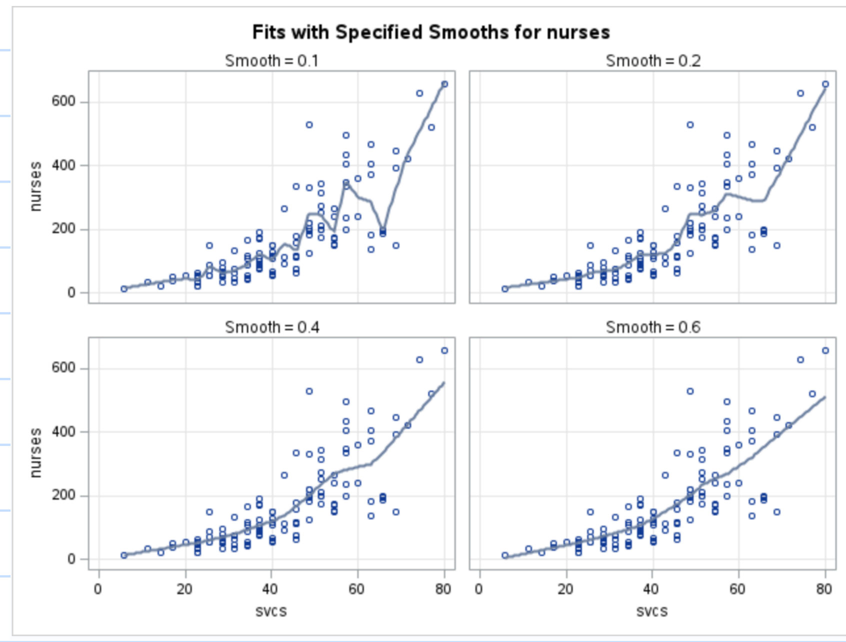


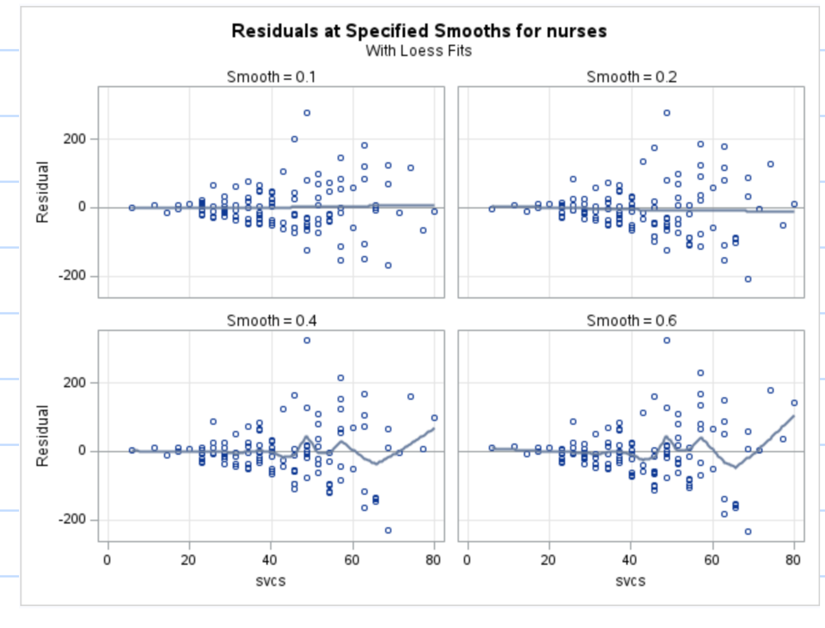
Model b:





Model c:



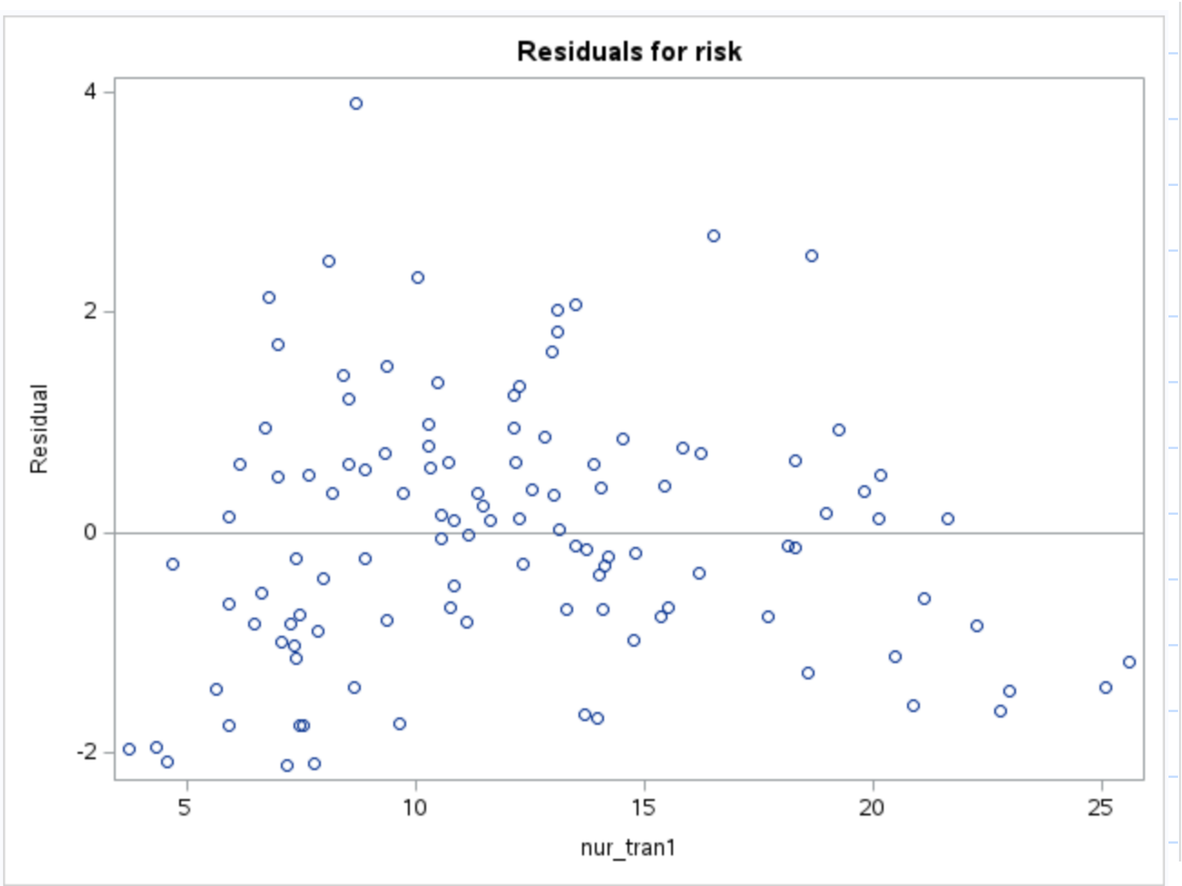


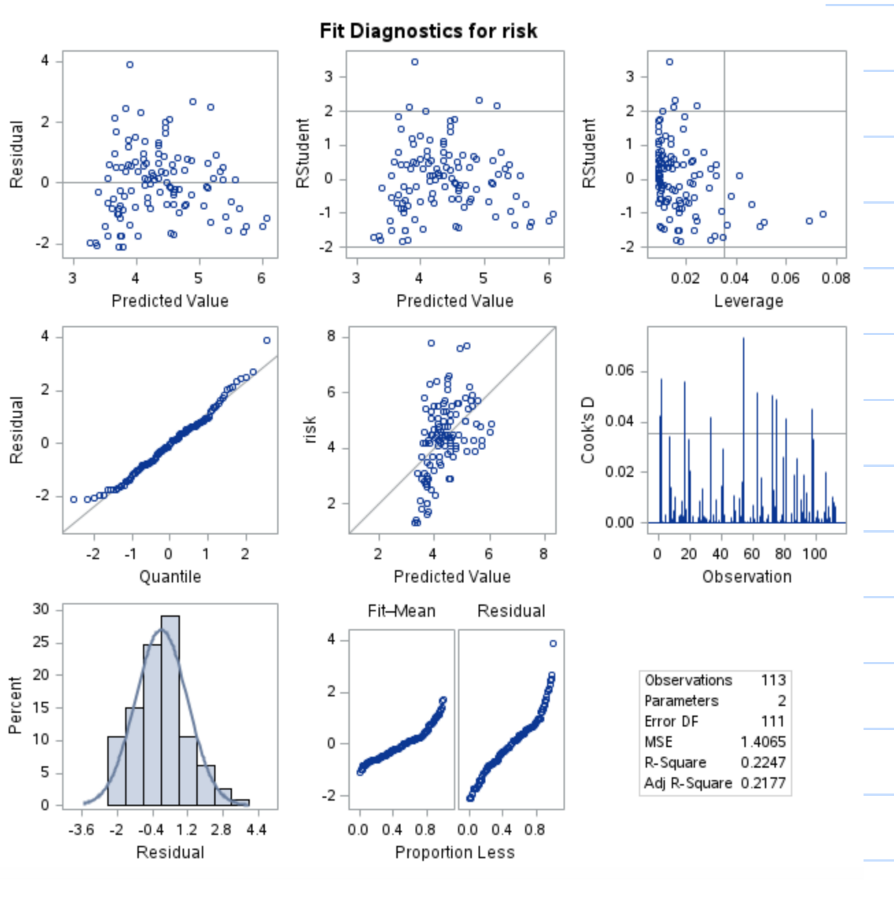
For model a, the loess fits in the plots confirm that 0.2 is a good smoothing value. Residuals for smoothing parameter value 0.1 and 0.2 exhibit less pattern (a "null" plot) than others. And the plot for smoothing parameter value 0.2 is smoother than that for 0.1.

For model b, the loess fits in the plots confirm that 0.2 is a good smoothing value. Residuals for smoothing parameter value 0.2 exhibit less pattern (a "null" plot) than others. And the plot for smoothing parameter value 0.2 is smoother than that for 0.1.

For model a, the loess fits in the plots confirm that 0.2 is a good smoothing value. Residuals for smoothing parameter value 0.1 and 0.2 exhibit less pattern (a "null" plot) than others. And the plot for smoothing parameter value 0.2 is smoother than that for 0.1.

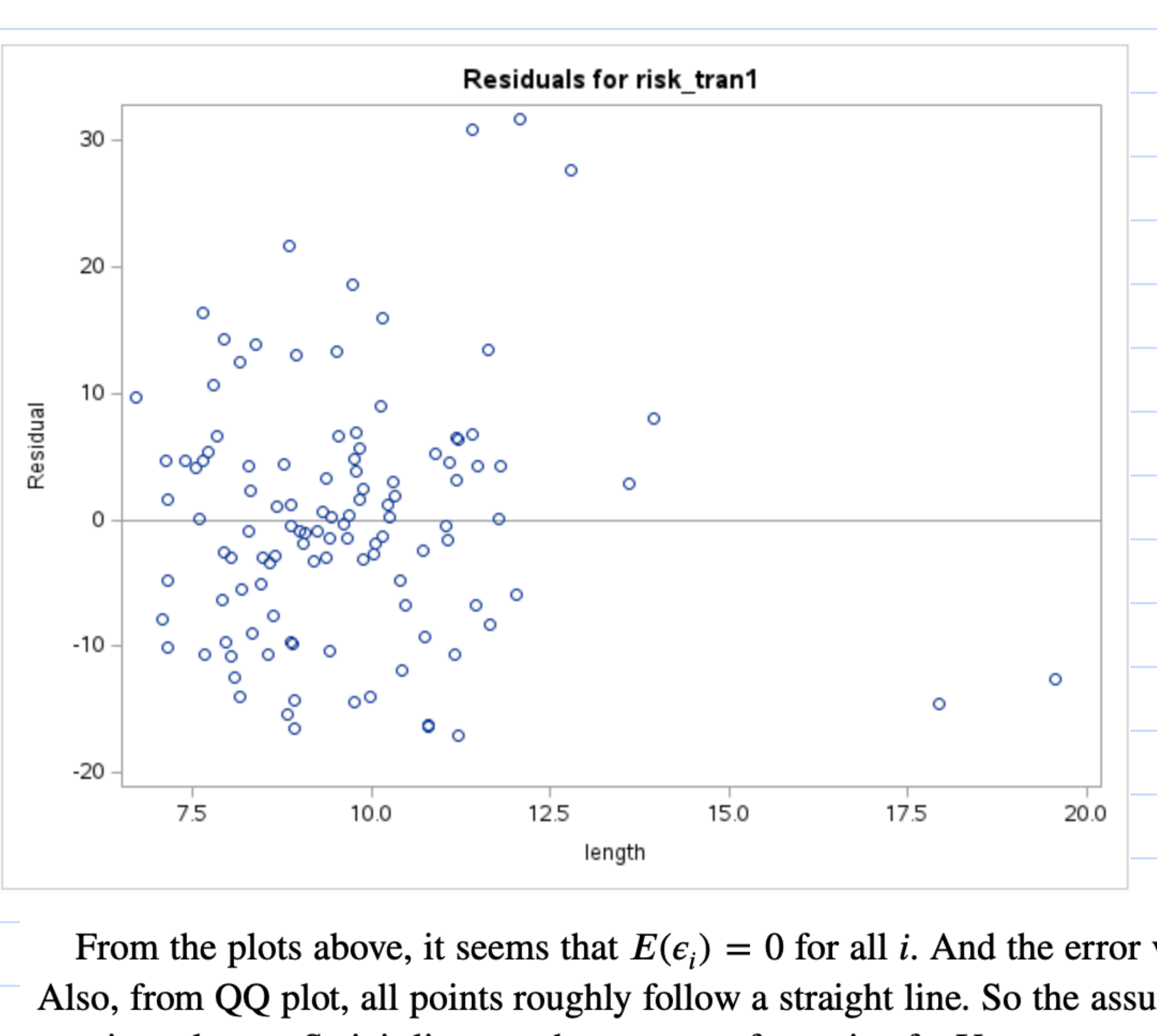
From the above plots, we notice three models are all nonlinear and all three models can become linear by some transformations.
For model a, we should make power transformation for Y or root transformation for X and here we take square root of X. The results are as follows:

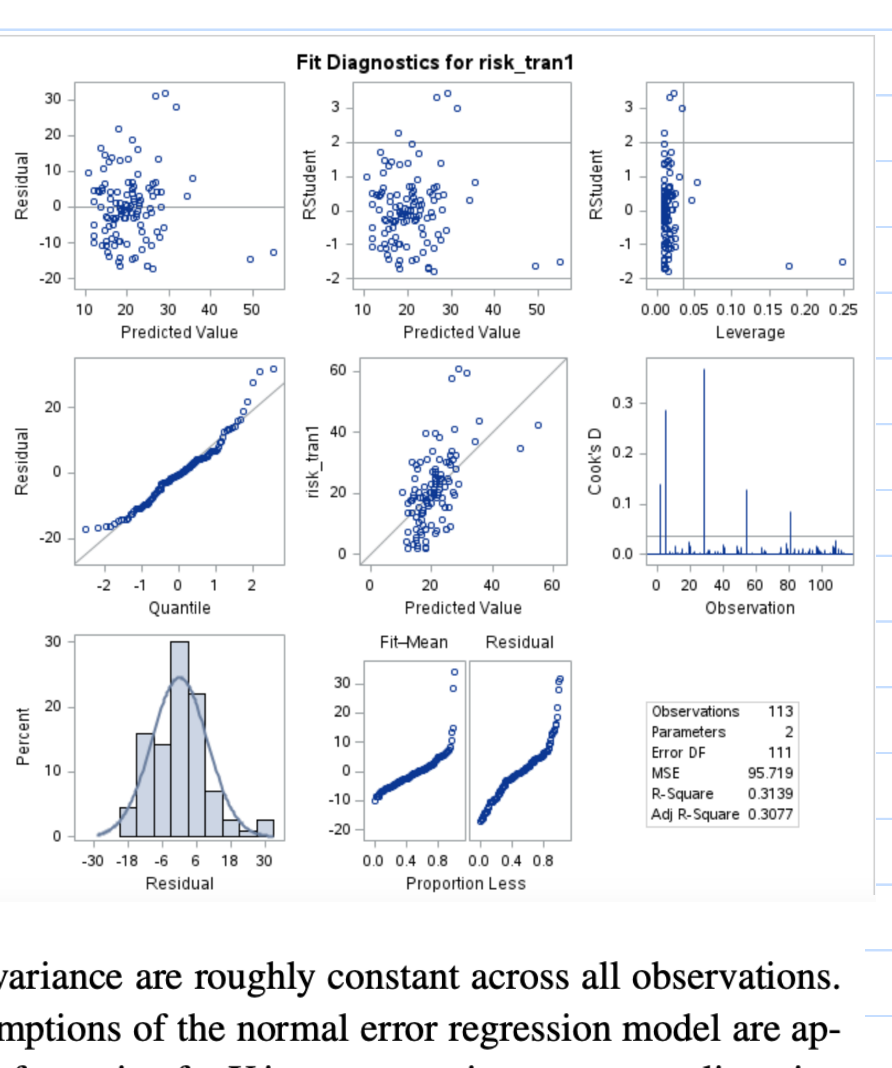




From the plots above, it seems that $E(\epsilon_i) = 0$ for all i. And the error variance are roughly constant across all observations. Also, from QQ plot, all points roughly follow a straight line. So the assumptions of the normal error regression model are approximately met. So it is linear and power transformation for Y or root transformation for X is an appropriate strategy to linearity.

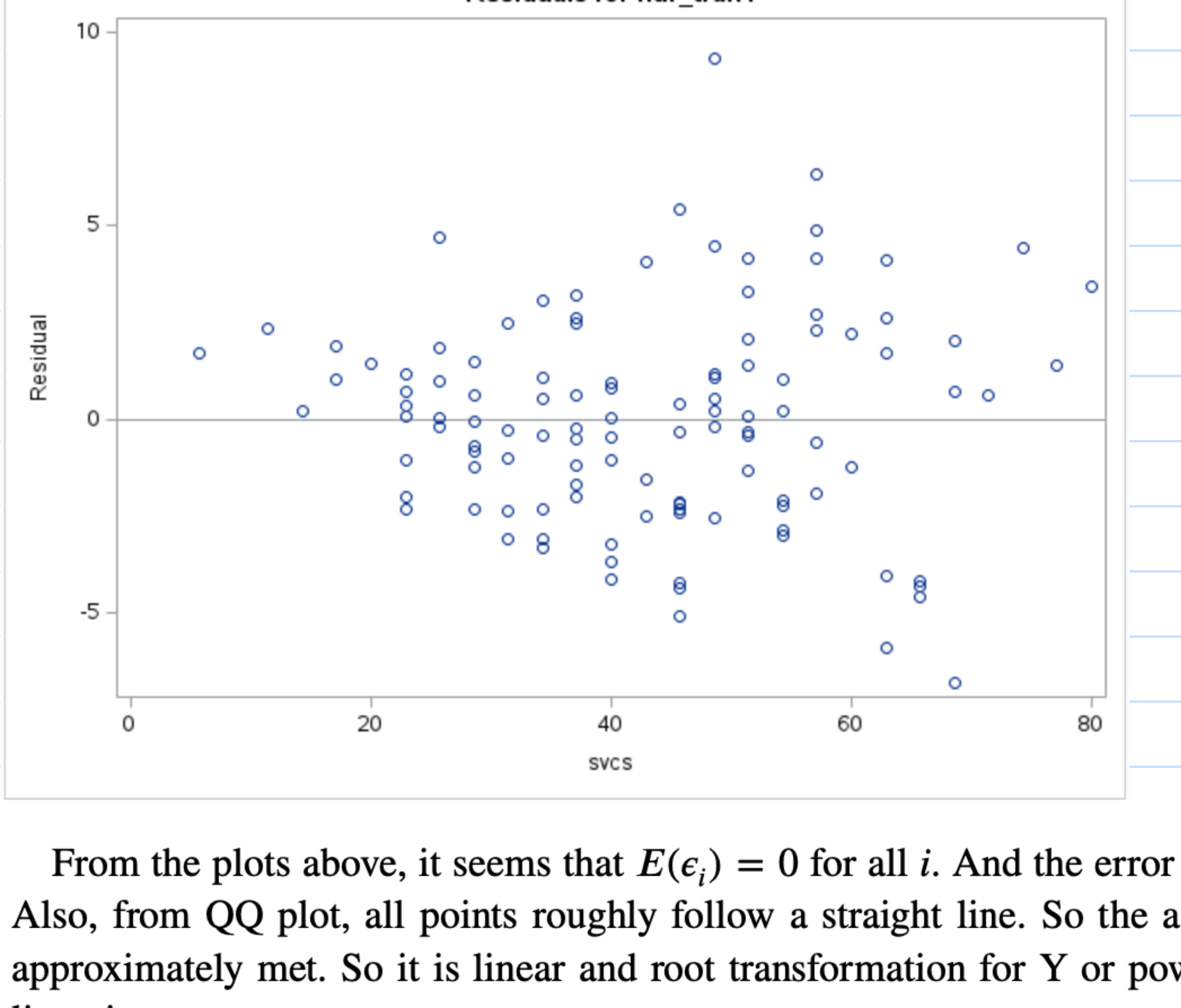
For model b, we should make power transformation for Y or root transformation for X and here we take square of Y. The results are as follows:

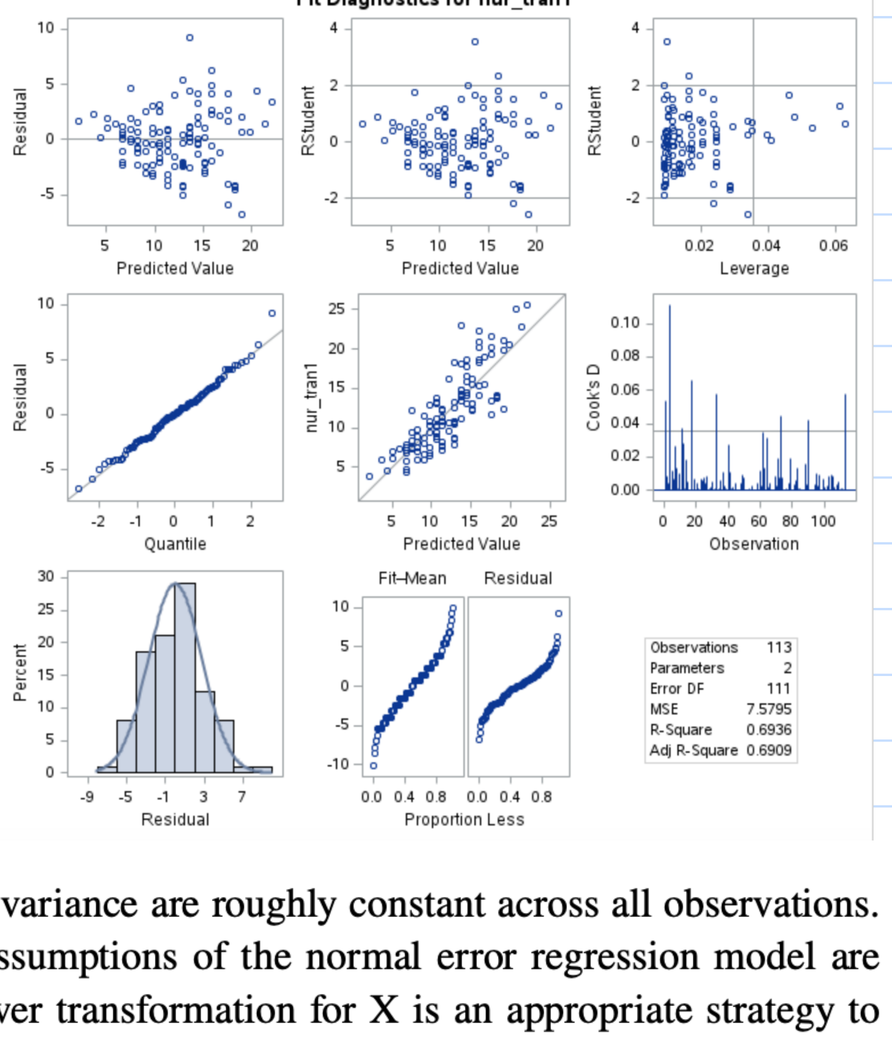




From the plots above, it seems that $E(\epsilon_i) = 0$ for all i. And the error variance are roughly constant across all observations. Also, from QQ plot, all points roughly follow a straight line. So the assumptions of the normal error regression model are approximately met. So it is linear and power transformation for Y or root transformation for X is an appropriate strategy to linearity.

For model c, we should make root transformation for Y or power transformation for X and here we take square root of Y. The results are as follows:





From the plots above, it seems that $E(\epsilon_i) = 0$ for all i. And the error variance are roughly constant across all observations. Also, from QQ plot, all points roughly follow a straight line. So the assumptions of the normal error regression model are approximately met. So it is linear and root transformation for Y or power transformation for X is an appropriate strategy to linearity.

4. For the simple linear regression model, show that (a) $\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$ and (b) $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

(1) $\sum(X_i-\bar{X})(Y_i-\bar{Y}) = \sum(X_i-\bar{X})Y_i - \sum(X_i-\bar{X})\bar{Y} = \sum(X_i-\bar{X})Y_i - \bar{Y}(\sum X_i - \sum\bar{X})$

Since $\sum X_i - \sum\bar{X} = 0$, then $\sum(X_i-\bar{X})(Y_i-\bar{Y}) = \sum(X_i-\bar{X})Y_i$.

So $$\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$$

(2) $Cov(\bar{Y}, \hat{\beta}_1) = Cov\left(\frac{\sum y_i}{n}, \sum c_i y_i\right) = \frac{1}{n} Cov(\sum y_i, \sum c_i y_i)$, where $c_i = \frac{x_i - \bar{x}}{\sum(x_i-\bar{x})^2}$

Since $Y_i \sim$ indep $N(\beta_0+\beta_1 x_i, \sigma^2)$, then $cov(y_i, y_j) = 0$ if $i \neq j$.

So $Cov(\bar{Y}, \hat{\beta}_1) = \frac{1}{n} Cov(\sum y_i, \sum c_i y_i) = \frac{1}{n}\sum c_i Var(y_i) = \frac{\sigma^2}{n}\sum c_i = \frac{\sigma^2}{n}\frac{\sum x_i - \sum\bar{x}}{\sum(x_i-\bar{x})^2} = 0$